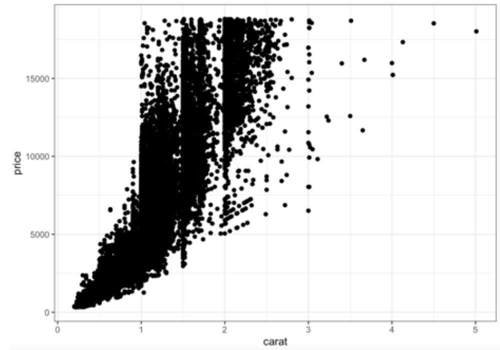
12.3 Data Exploration – Overview

Scatterplots

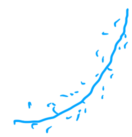


- Displays the relationship between **two quantitative** variables measured on the same individuals.

- Useful to determine if an **association** exists!
  - So is there a pattern where *some values of one variable* tend to occur with *some values of the other variable*.
  - Ex) Smaller carat diamonds tend to have lower prices, and as the carat increases prices tend to increase as well.

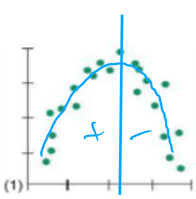
- Setup of axes
  - The explanatory (independent) variable goes on the X (horizontal) axis.
  - The response (dependent) variable goes on the Y (vertical) axis.
  - Ex) How large a diamond is impacts how much it costs → Carat = *X*; Price = *Y*.

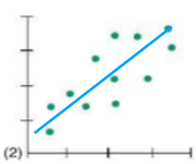
- Interpreting a scatterplot (what we are looking for in a scatterplot)
  - **Form** (pattern of the dots)
    - Linear → Points follow a general linear trend; Straight line.
    - Curved → Points show some evidence of curvature; NOT a straight line.
    - Random scatter → No pattern, points are just scattered about randomly kinda like a cloud of points.

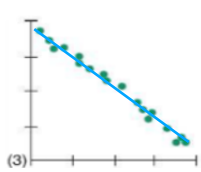


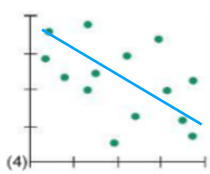
  - **Direction** (of the association; only applies to linear relationships)
    - Positive → Upward trend.
    - Negative → Downward trend.
    - No Association →There is no pattern or general trend (corresponds to random scatter).

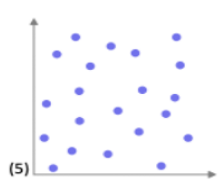
  - **Strength** (how strong the association is; how well the data fits the pattern; only applying this to linear relationships)











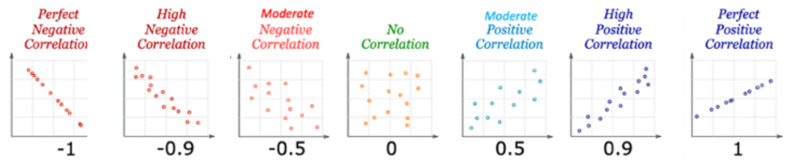
Example

	Form	Direction	Strength
(1)	Curved	N/A (+/-)	N/A (strong)
(2)	Linear	Positive	Moderate
(3)	Linear	Negative	Strong
(4)	Roughly Linear	Negative	Weak
(5)	Random Scatter	No association	Very Weak

Correlation

- The **correlation (*r*)** is an index that expresses the direction and strength of the relationship.
  - It combines both of these aspects into a single number measure.
  - Often referred to as the correlation coefficient (or Pearson's correlation).

- Interpreting correlation
  - Sign = Direction
  - Absolute value $|r|$ = Strength



- Properties of Correlation
  - Scale goes from -1 to 1 → $-1 \le r \le 1$
  - Only applies to LINEAR relationships.
  - *r* has no units and is the same regardless of which variable is *X* or *Y*.
  - Does NOT imply a cause-and-effect relationship.

    Ex) Ice cream sales and shark attacks have a strong positive correlation.

Using your Calculator!

Using TI-83/84 (and TI-30 XS MultiView / XIIS) to calculate correlation (and regression line).

Steps for the TI-83/84	Steps for the TI-30XS MultiView	Steps for the TI-30 XIIS
1. Enter data: STAT → Edit → Enter X data in L_1 Enter Y data in L_2 2. Calculate: STAT → CALC → LinReg(ax+b) a) XList: L_1. b) YList: L_2. c) Rest leave blank. d) Calculate!	1. Data → Enter X data in L_1 Enter Y data in L_2 2. 2^{nd} → stat → 2-Var Stats a) *x*DATA: L1 b) *y*DATA: L2 c) CALC	1. 2^{nd} → STAT → 2-VAR (Enter) 2. DATA X_1 = # (scroll down) Y_1 = # (scroll down) … (repeat for all data points) 3. STATVAR (scroll across) 4. To exit this menu: 2^{nd} → EXIT STAT → Y

****** One time setup for TI-83/84: 2^{nd} → Catalog → DiagnosticOn → Enter*

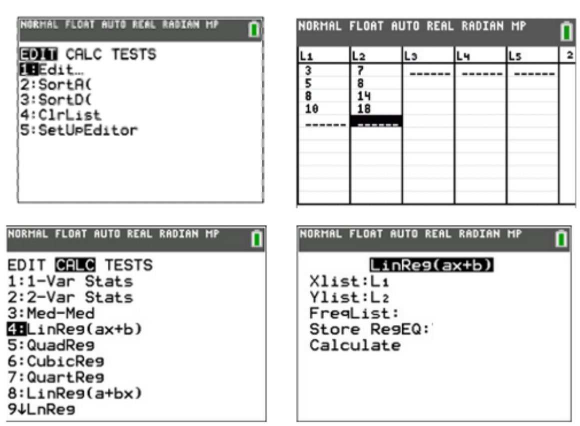
Demo dataset

X	Y
3	7
5	8
8	14
10	18

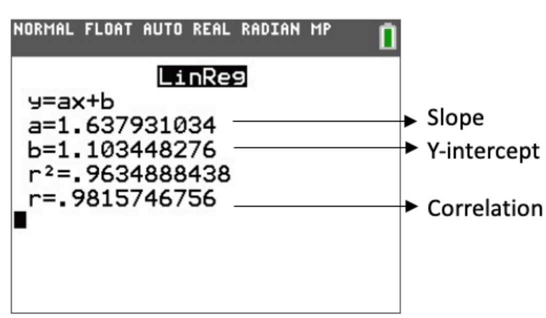
n=4

L_1 L_2

Inputs



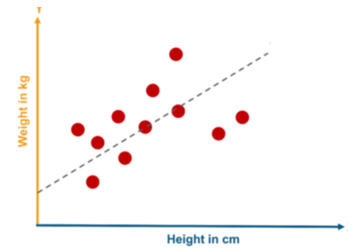
Results



Regression

- Ultimately, we want to determine if we can use a straight line to model the relationship between two variables → If so, we can use that model to make predictions!
    - This process is called **Linear Regression**.

Step 1 → Determine if there is a **significant correlation** (**linear relationship**).



a) Calculate correlation.
b) Compare it to the **Table of Critical Values or the Pearson Correlation Coefficient** to see if it is statistically significant.
    - Match the sample size n and the Level of significance α (Probability our claims about the data are wrong) to the specific problem.
    - If $|r|$ > Critical Value (CV) → r is statistically significant (unlikely to have occurred by chance).

Critical Values of the Pearson Correlation Coefficient

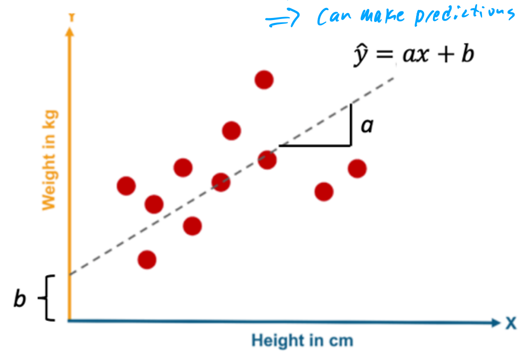
n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

Step 2 → Once we have a significant correlation, we can find the **regression line**.

- Linear equation that fits our data best (aka 'line of best fit').
    - It is IMPORTANT to get the X and Y variables correct!
- Our calculator gives us our equation!

Demo ex) $n = 4$, $\alpha = 0.05$

$r = 0.982 > 0.950 = CV$
⟹ significant ✓
⟹ Can make predictions ✓



Equation

- Here is the form of our linear equation (written in slope-intercept form):

$$\hat{y} = ax + b$$

- x = Value of the explanatory variable
- $\hat{y}$ = Predicted value of the response variable for the given x
- a = Slope
    - It measures the direction and steepness of the line
- b = Y intercept
    - It is the location where the regression line crosses the Y-axis (value of Y when X = 0)

Step 3 → Make **predictions** using the regression line.

- We can think of our regression line, and specifically $\hat{y}$, as predicted values of Y for all X values *in the X range of our sample data*!

- Calculating these is simple:
    - Just plug in the new X value to our equation and this will give us the predicted Y.

    - Demo example) Predict Y for X = 7.

$$\underbrace{1.638x}_{ax} + \underbrace{1.103}_{b} \quad \hat{y} = 1.638x + 1.103 \longrightarrow \hat{y} = 1.638(7) + 1.103 = 12.569$$

Full Example

Top Row X **Hours Spent on Homework**	41	20	34	43	9	20	54	52	10	21
Y **Grade on Test**	79	63	76	100	55	82	95	80	60	80

Part 1 Part 2

a) Calculate the correlation for the dataset above and determine if it is statistically significant at a level of significance of $\alpha = 0.05$.

Show work: → linreg (ax+b)
X = HW, Y = Grade

OR

→ 2 var stats (X = HW, Y = Grade)

Part 1 → Calculate
$\boxed{r = 0.779}$

Part 2 → Comparison to CV
$n = 10$, $\alpha = 0.05$

$r = 0.779 > 0.632 = CV$

⇒ significant ✓

b) If appropriate, determine the regression equation.

↳ yes because r is significant

$\hat{y} = ax + b$
$\boxed{\hat{y} = 0.676x + 56.459}$

c) If a student spends 35 hours on homework, make a prediction for their grade on the test.

$\hat{y} = 0.676(35) + 56.459$
$\boxed{\hat{y} = 80.119}$

d) If a student spends 50 hours on homework, make a prediction for their grade on the test.

$\hat{y} = 0.676(50) + 56.459$
$\boxed{\hat{y} = 90.259}$